I decided to use Python Pandas and MySQL for the data transformation because these data tools are primarily used in their tech stack and I would be able to perform these data transformations within their own server.

Note: I used Python Pandas to perform cleaning, formatting, and calculating new metrics from legacy data before uploading data to query on MySQL Server. I also merged data in Python script to create a Staging Table. The results of the query from MySQL Server exports the data in the company’s usable format which can then be uploaded directly to the company’s pipeline. The resulting import files are found in the “EXPORT\_BIN” folder.

The end user can then use the transaction data in MySQL to compute their own return; they can do this easily on all 3 reporting levels (client, household, & account) by using the computed daily prices of each security along with the historical transaction data on each reporting level. I used MySQL to perform the transformations and then read the results in Python to export the results in the appropriate import template via a CSV file.

Step-by-Step Coding Procedure:

-Data Cleaning

-Data Staging in Pandas table

* Derive price-per-share per day from division of account value and quantities of securities by iterating through each file and dividing the numbers from each day

-Make connection to MySQL Database

* Upload all legacy data into appropriate tables

-Perform data transformation on legacy data, via querying, to match the “Historical Prices” and “Transactions” import templates

* Convert Transaction types anf Securtity types from legacy codes to new codes using mapping files found in “ReferenceFiles” folder

-Export resulting data to CSV in “EXPORT\_BIN” folder

Closing thoughts:

All in all, this Case Study was a very challenging yet rewarding experience in which I had to consider a real-world case of data cleaning and data manipulation to help onboard a client from their legacy system and into the new system while following an import template that allows the company to dynamically produce financial reports on multiple reporting levels. I am more confident in my ability to transform data to its required format and was forced to think critically about the performance and efficiency of my program as it would be run on massive scale of data. I thoroughly enjoyed this challenge, learning from this opportunity and am happy to be one step closer to my goal of providing robust data solutions.